



# L'analisi dei testi

Stefano Ballerio

# Agenda

- Il sovraccarico informativo e la knowledge discovery
- Informazioni, conoscenze e metodi informatici
- Le applicazioni: alcuni esempi
- Il text mining
  - Fasi del text mining
  - Il text preprocessing
  - Le basi di conoscenze
  - Il mining
  - L'interpretazione dei risultati
- La ricerca in corso: lettere di dimissione e text mining
- Bibliografia

# Il problema: il sovraccarico informativo e la necessità di conoscenze

«The volume of published **biomedical research**, and therefore the underlying biomedical **knowledge base**, is **expanding** at an increasing rate». Cohen e Hersh (2005)

PubMed:

- abstract di oltre 12.000.000 di research papers
- 40.000 nuovi abstract al mese

Da queste basi di dati è necessario **estrarre conoscenze** in modo efficiente.

# La soluzione: la knowledge discovery e il text mining

«Among the tools that can aid researchers in coping with this information overload are **text mining** and **knowledge extraction**».  
Cohen e Hersh (2005)



# Data mining e knowledge discovery in database

Analisi dei dati strutturati, di tipo quantitativo e numerico ➡ tecniche di **data mining** (DM)

DM integra metodi statistici e informatici per estrarre nuove conoscenze da grandi basi di dati strutturati.

Si parla di **knowledge discovery in database** (KDD).

# Text mining e knowledge discovery in text

La maggior parte delle informazioni disponibili negli archivi delle istituzioni scientifiche e delle società commerciali è però nella forma di testi scritti in linguaggio naturale.



Servono strumenti per l'estrazione di conoscenze da grandi collezioni di testi in linguaggio naturale: il **text mining** (TM) e la **knowledge discovery in text** (KDT).

# Knowledge discovery in text

La KDT è «il processo non banale di **identificazione di pattern** validi, nuovi, potenzialmente utili e infine comprensibili **in dati testuali**». Dulli, Polpettini e Trotta (2004)

- **valido**: il pattern è proiettabile su nuove collezioni di testi
- **nuovo**: il pattern era precedentemente sconosciuto a colui che lo ha individuato
- **utile**: il pattern consente di confutare o corroborare un'ipotesi o può essere usato per supportare scelte strategiche e processi decisionali
- **comprensibile**: per dimensioni, complessità e rappresentazione, il pattern può essere compreso dall'utente umano



# Text mining

«Text mining can be broadly defined as a knowledge-intensive process in which a **user** interacts with a **document collection** over time by using a suite of **analysis tools**. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the **identification and exploration of interesting patterns**. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections». Feldman e Sanger (2007)



# Tecniche e metodi affini

La KDT implica di norma il ricorso a tecniche e metodi affini:

- **natural language processing** (NLP)
- **information retrieval** (IR)
- **information extraction** (IE).

La KDT richiede infatti di individuare i materiali testuali necessari, trattarli automaticamente ed estrarne informazioni.

# Applicazioni (1/2)

- **Corporate finance e business intelligence**

tendenze e associazioni ricorrenti in relazione a transazioni, compagnie, prodotti, persone

- In ambito biomedico, **classificazione della letteratura scientifica**

costruzione di database annotati di articoli, abstract, relazioni ecc.

- **Hypothesis generation in complementary structures in disjoint literatures**

TM su n articoli: A influisce su B  
TM su m articoli: B influisce su C  
allora (ipotesi) A influisce su C

Swanson (1991) correla così emicrania e carenza di magnesio.

# Applicazioni (2/2)

- Database di referti radiologici + TM = conoscenze epidemiologiche sui pazienti sottoposti a esame radiografico

Fizman, Chapman, Aronsky, Evans e Haug (2000)

Hripcsak, Austin, Alderson e Friedman (2002)

- Lettere di dimissione + TM = rilevamento delle complicanze

Melton e Hripcsak (2005) applicano TM a un database di lettere di dimissione e rilevano le complicanze definite dal New York Patient Occurrence Reporting and Tracking System.

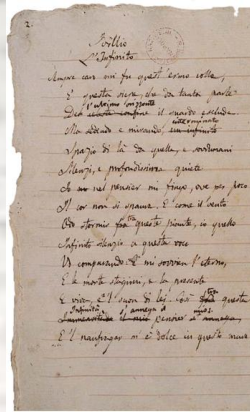


# Fasi del text mining

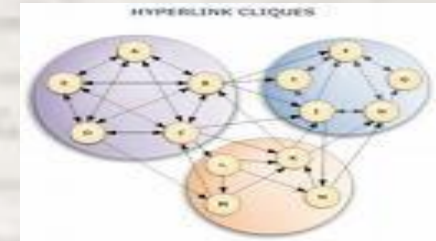
➤ text preprocessing



➤ costituzione del corpus



➤ mining



➤ interpretazione e valutazione dei risultati



➤ obiettivi della ricerca  
➤ euristica: concetti e relazioni  
➤ traduzione dell'euristica in termini linguistici e statistici





# Il text preprocessing: obiettivo

- Obiettivo del text preprocessing è **trasformare ogni testo in una rappresentazione esplicitamente strutturata**.
- Quali **tratti** dei testi sono **rilevanti**?  
Dipende dagli obiettivi e all'euristica definiti.
- Il preprocessing è molto più complesso nel TM che nel DM: un testo è un oggetto strutturato (morfologia, sintassi, punteggiatura, testualità, redazione, grafica), ma lo è in modo complesso e non evidente.

# Tratti (1/2)

I tratti (features) usati per rappresentare i testi sono caratteri, parole, termini e concetti:

- **caratteri**
  - lettere, numeri e altri caratteri speciali
  - si possono cercare o eliminare (!, \*, >)
- **parole**
  - ogni parola singola
  - si possono cercare solo alcune parole (entry list) o se ne possono eliminare (stop list: tipicamente, le parole funzione come gli articoli)
- **termini**
  - parole singole (*miocardio*) o espressioni composte (*Casa Bianca*, *struttura ospedaliera*)
  - possono essere ricondotti a termini normalizzati

# Tratti (2/2)

- **concetti**
  - a uno stesso concetto (*astensione dall'alcol*) si riconducono diverse espressioni
  - il concetto include non solo varianti ortografiche (*astensione dall'alcol* o *astensione dall'alcool*) e morfologiche (*astenersi dall'alcol*), ma anche sinonimi (*astenersi da bevande alcoliche*) ecc.

Come rappresentare  
la presenza/assenza  
di un tratto?



a) con un'opposizione binaria 0/1



b) con una pesatura, sulla base  
dei valori di term frequency e  
inverse document frequency

# Termini e concetti

- I risultati migliori si ottengono con termini e concetti .
- Per estrarli, spesso si deve ricorrere a basi di conoscenze come **ontologie** e **dizionari**.
- Problemi con i concetti:
  - **quali** sono i concetti interessanti?
  - richiedono **lavoro umano**
  - sono **legati a un dominio** e quindi non sono facilmente esportabili.



# Livelli di analisi e di descrizione del testo

- Livello **lessicale**: parole e termini
- Livello **sintattico**: sintagmi
- Livello **semantico**: coreferenza, concetti e relazioni

# Il text preprocessing come information extraction

- Il text preprocessing si struttura come un processo di **information extraction**.
- Possiamo estrarre elementi diversi:
  - **entità**: persone, patologie, geni o proteine
  - **attributi**: l'età o il ruolo di una persona
  - **fatti**: la sussistenza di una relazione tra un gene e una proteina
  - **eventi**: un episodio di scompenso o l'impianto di un by-pass.

# Compiti di IE nel preprocessing (1/2)

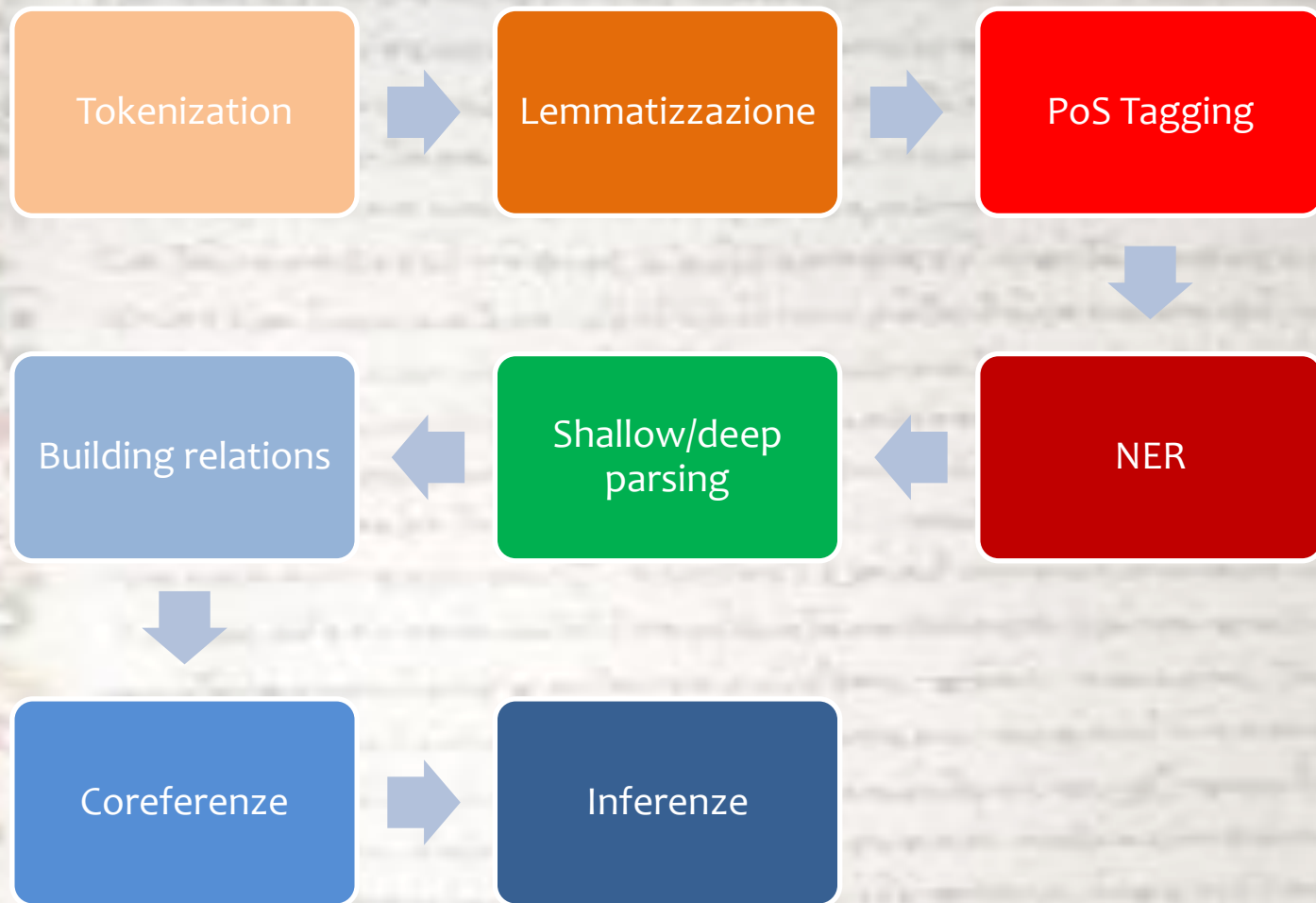
- **Template element**: riconoscere persone, organizzazioni, luoghi e artefatti non dominio-specifici
  - nome = Ippocrate // categoria = persona
  - nome = A. Osp. di Melegnano //
  - categoria = organizzazione
- **Template relationship**: riconoscere relazioni tra gli elementi individuati
  - lavora\_presso[N-Y]
  - lavora\_presso[Ippocrate-A. Osp. di Melegnano]

# Compiti di IE nel preprocessing (2/2)

- **Scenario templates**: riconoscere elementi e relazioni specifici di un dominio
  - attiva[proteinaX-geneY]
- **Coreferenza**: individuare le catene anaforiche e cataforiche formate da nomi, pronomi e altre espressioni
  - Il **paziente** si è ripresentato e il medico **gli** ha detto di curarsi



# Information flow di un sistema di IE (1/3)



# Information flow di un sistema di IE (2/3)

## 1. Tokenization

1. si scompone il testo in unità o **token**, ovvero in parole e poi in periodi e capoversi
2. in questa fase si possono eliminare le **stop words**
3. esempio: # paziente // lamenta // ~~un~~ dolore // acuto // ~~al~~ torace

## 2. Lemmatizzazione

1. ogni parola è ricondotta al lemma relativo
2. esempio: dolore, dolori → dolore

## 3. Part of speech tagging

1. per ogni parola si indica la parte del discorso che essa esemplifica (nome, aggettivo, verbo ecc.)
2. esempio: [dolore – N], [lamentare – V]

## 4. Proper name identification o named entity recognition (NER)

1. la NER interessa nomi propri di persone, organizzazioni e luoghi; date e indicazioni temporali; valute, percentuali, numeri di telefono ecc.
2. esempi: Ippocrate, 4 luglio p.v., 44%

# Information flow di un sistema di IE (3/3)

## 5. **Shallow parsing**

1. sulla base delle informazioni estratte finora e di un insieme di pattern ricorrenti definiti manualmente, il sistema individua i sintagmi nominali e i sintagmi verbali
2. esempi: [*il paziente scompensato* – SN], [*si sente male* – SV]

## 6. **Building relations**

1. le relazioni tra le entità individuate sono costruite attraverso pattern specifici
2. esempio: [persona] [dichiara] [sintomo]
3. ogni elemento del pattern è un'etichetta per un insieme di forme linguistiche:  
[persona] = {paziente, soggetto, la signora X, il signor X... }

## 7. **Coreferenze**

1. si individuano le relazioni di coreferenza

## 8. **Inferenze**

1. il sistema può includere regole di inferenza per estrarre ulteriori informazioni
2. esempio: se il testo dice che Marco vive con Anna (relazione) e che Marco abita a Milano (relazione), allora si può inferire che Anna abita a Milano  
la regola necessaria è: se AB[vive con] e AX[abita in], allora BX[abita in]


# Difficoltà (1/2)

## Feature dimensionality

Troppi tratti significativi!

## Feature sparsity

Molte parole solo in pochi testi!

- 
1. Aumenta il **costo computazionale**
  2. Sovrabbondanza di pattern: i risultati sono di **difficile lettura**



# Difficoltà (2/2)

La rappresentazione del testo deve unire **accuratezza informativa** ed **efficienza computazionale**:

- ridurre le parole, eliminando le stop words
- ridurre i tratti, selezionando quelli importanti e discriminanti
- raffinare le tecniche da applicare alle risposte.

# Ontologie e dizionari

- Il processo di IE richiede spesso il ricorso a **ontologie** e **dizionari**.
- Ontologia: sistema strutturato di concetti relativi al dominio in esame.
- Quando si lavora su un dominio semantico per uno scopo specifico, le basi di conoscenze sono decisive.

# Quali basi?

- **Standard** o costruite **ad hoc**
- **Generiche** (dizionari generici, WordNet) o **settoriali**
- Per il settore biomedico: la versione più recente di **UMLS** (2007AA) integra 139 sorgenti terminologiche e ne trae un sistema concettuale codificato.
- La scelta delle basi di conoscenza è sempre connessa all'applicazione per la quale si costruisce il sistema e quindi al dominio in esame.
- E **in Italia**? Poche risorse disponibili

# Uso delle ontologie

- Le ontologie possono intervenire in momenti diversi



- **Uso attivo** delle ontologie: i concetti dell'ontologia informano la ricerca.



# Il mining

«The core functionality of a text mining system resides in the analysis of concept co-occurrence patterns across documents in a collection». Feldman e Sanger (2007)

# Che cosa si cerca

**Pattern** che si cercano nella fase di mining:

- **distribuzioni** di concetti e relazioni
- **frequent sets** e **near frequent sets** di testi per concetti e relazioni
- **associazioni** tra concetti e relazioni.

# L'analisi di un corpus nel tempo

- **Analisi dei trend:** confronto tra sottoinsiemi cronologicamente definiti del corpus
- **Algoritmi incrementali:** aggiornamento progressivo dei risultati dell'analisi del corpus

# Text categorization

- La **text categorization** (TC) è la classificazione dei testi di un corpus in una serie di categorie predefinite.
- Usi di TC:
  - indicizzazione di testi sulla base di un vocabolario controllato
  - document sorting
  - filtraggio.
- Aspetti di TC:
  - **single-label** (ogni testo appartiene a una sola categoria) o **multilabel** (ogni testo può appartenere anche a più di una categoria)
  - per **decisioni binarie** o per **pesature**.



# Clustering

- Il **clustering** è il raggruppamento dei testi di un corpus in categorie non predefinite, sulla base della **similarità** tra i testi stessi.
- Ipotesi di base: i testi rilevanti si assomigliano tra loro più che con quelli non rilevanti.
- Il clustering può essere **gerarchico** (in ordine di similarità) o **ricorsivo** (ovvero uno *scatter-gather*).

# L'interpretazione dei risultati

- L'interpretazione dei risultati avviene alla luce dell'**euristica** e degli **obiettivi** iniziali.
- Per agevolare l'interpretazione si possono raffinare i metodi di interrogazione del sistema e di visualizzazione dei risultati (GUIs)

# Fasi del text mining - riepilogo

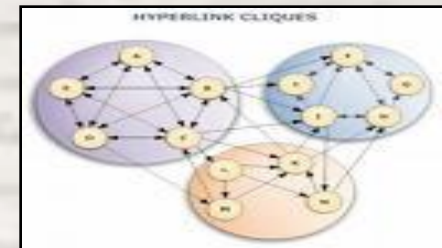
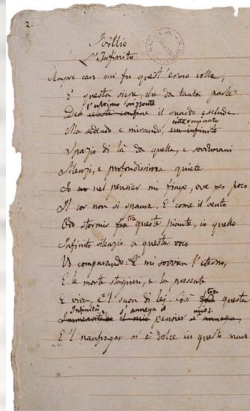
➤ text preprocessing

➤ costituzione del corpus

➤ mining

- obiettivi della ricerca
- euristica: concetti e relazioni
- traduzione dell'euristica in termini linguistici e statistici

➤ interpretazione e valutazione dei risultati



# La ricerca in corso: gli obiettivi

- Valutare la **continuità assistenziale** limitatamente alla sua **dimensione informazionale**
- Sperimentare degli **indicatori** della continuità assistenziale
- Sviluppare e applicare **strumenti di analisi automatizzata dei testi**

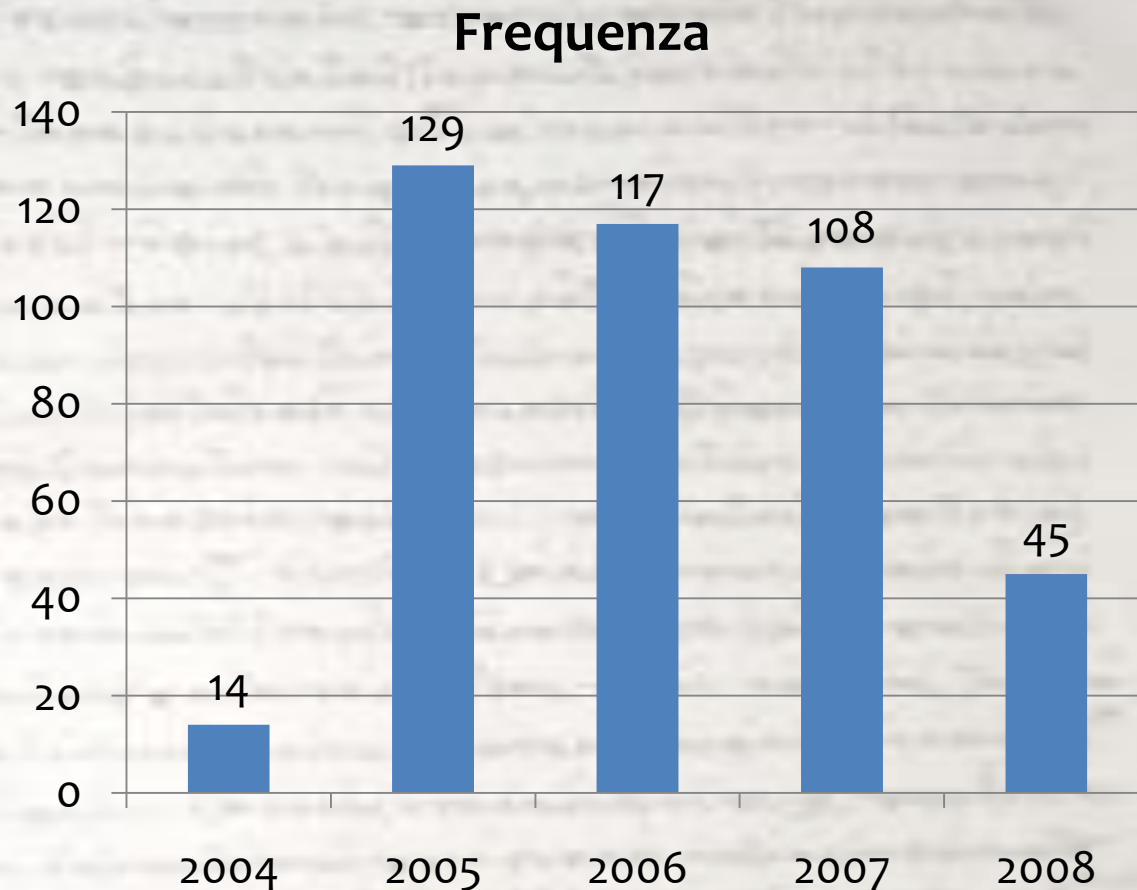


# Il modello della Joint Commission

- **Accesso e continuità dell'assistenza (ACC)**
- **Standard ACC.3.2:** la cartella clinica contiene una copia della lettera di dimissione
- **Intento di ACC.3.2:** al momento della dimissione viene stilata una relazione delle cure fornite al paziente. La lettera di dimissione viene compilata da un soggetto qualificato, che può essere il medico che ha in carico il paziente, un altro medico ospedaliero (es. medico di guardia) o un impiegato.
- La lettera di dimissione contiene:
  - Il motivo del ricovero
  - Riscontri e accertamenti fisici e di altro genere significativi
  - Diagnosi e comorbilità significative
  - Procedure diagnostiche e terapeutiche eseguite
  - Terapia farmacologica significativa e altre terapie significative
  - **Condizioni del paziente alla dimissione**
  - **Terapia farmacologica alla dimissione**, tutti i farmaci da assumere al domicilio
  - **Istruzioni di follow-up.**

# I test (1/2)

413 lettere di dimissione del reparto di cardiologia dell'ospedale Uboldo di Cernusco sul Naviglio. Le lettere sono relative a pazienti con scompenso cardiaco. L'intervallo temporale va dal 2004 al 2008.



# I test (2/2)

- I casi di scompenso cardiaco sono stati selezionati dal file delle **SDO** utilizzando i criteri **AHRQ** di ospedalizzazione evitabile per **scompenso cardiaco**.
- Dal file delle SDO si è risaliti alle **lettere di dimissione** tramite il numero di cartella clinica.
- Le lettere di dimissione sono in **formato Access<sup>TM</sup>** e contengono quattro campi memo, relativi ad anamnesi, esami effettuati, decorso e prescrizioni alla dimissione.

# Il confronto tra i testi e il modello

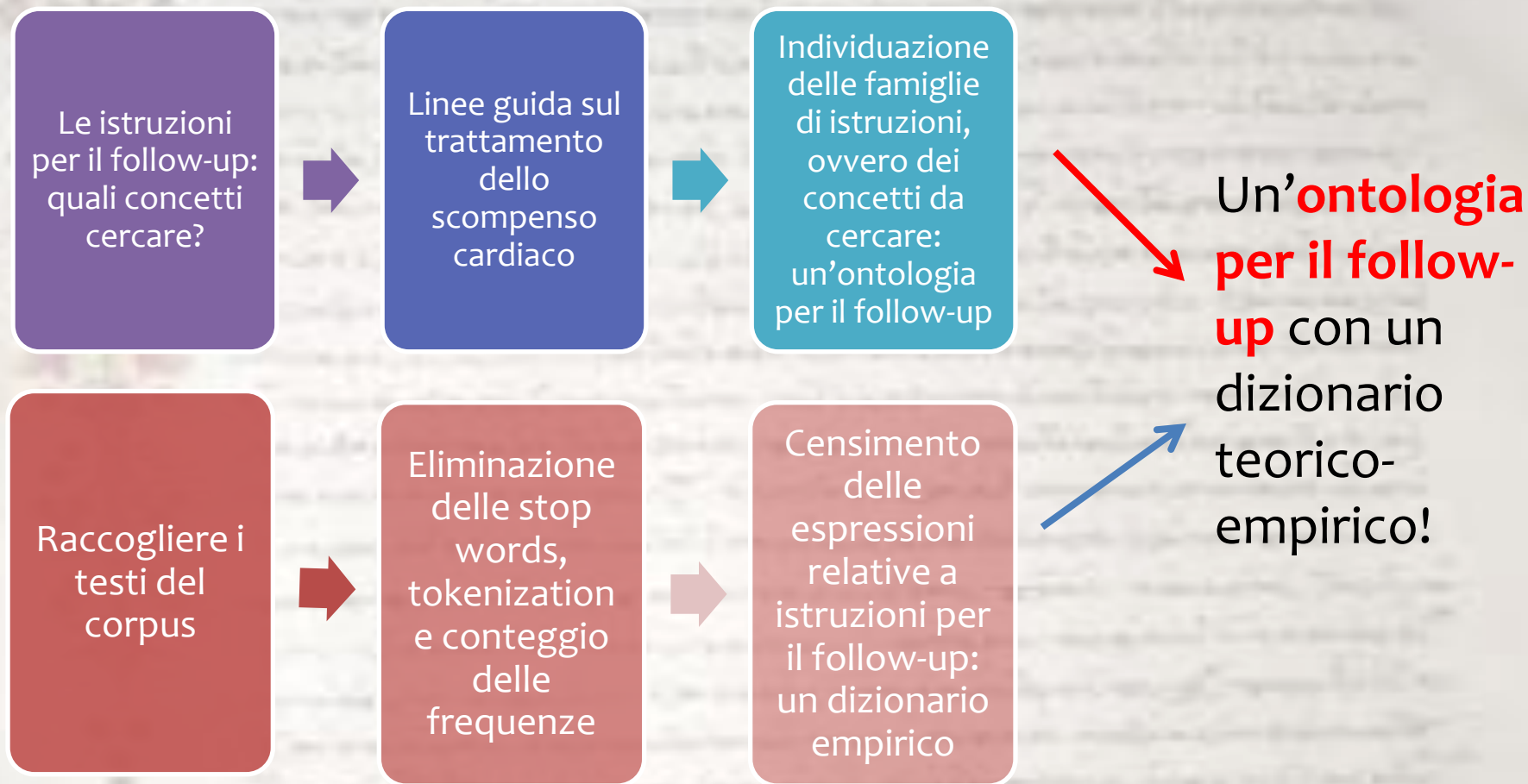
Creazione di  
un'ontologia per le  
istruzioni di follow-  
up

Esame dei testi del  
corpus alla luce  
dell'ontologia

Confronto tra i testi  
esaminati e il  
modello e  
valutazione della  
distanza



# Un'ontologia per il follow-up

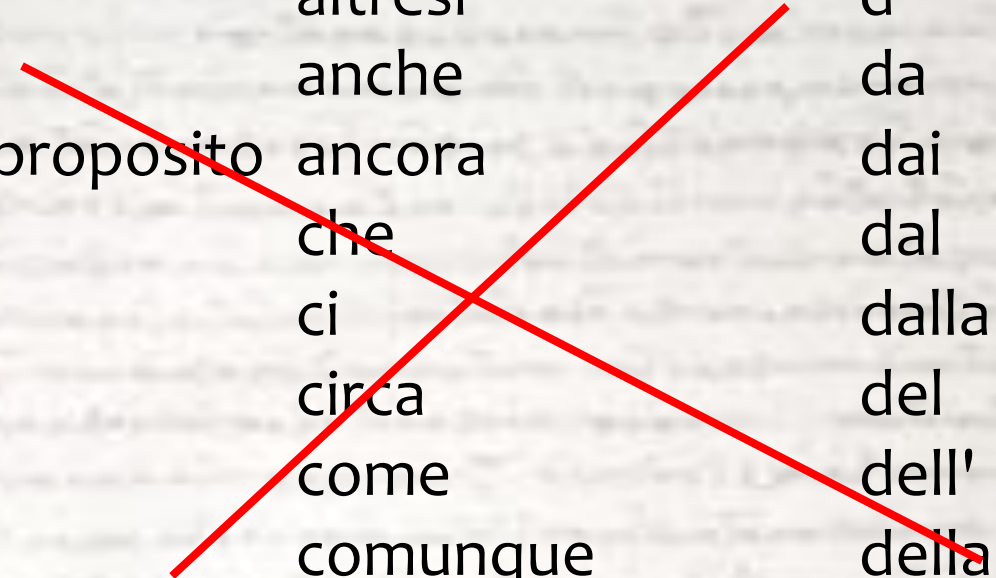


# Istruzioni per il follow-up: i concetti



# I testi del corpus: le stop words

a	allo	cui
a causa	altresì	d'
a fronte	anche	da
a questo proposito	ancora	dai
ad	che	dal
agli	ci	dalla
ai	circa	del
al	come	dell'
all'	comunque	della
alla	con	delle
alle	così	di...



# Tokenization e frequenze

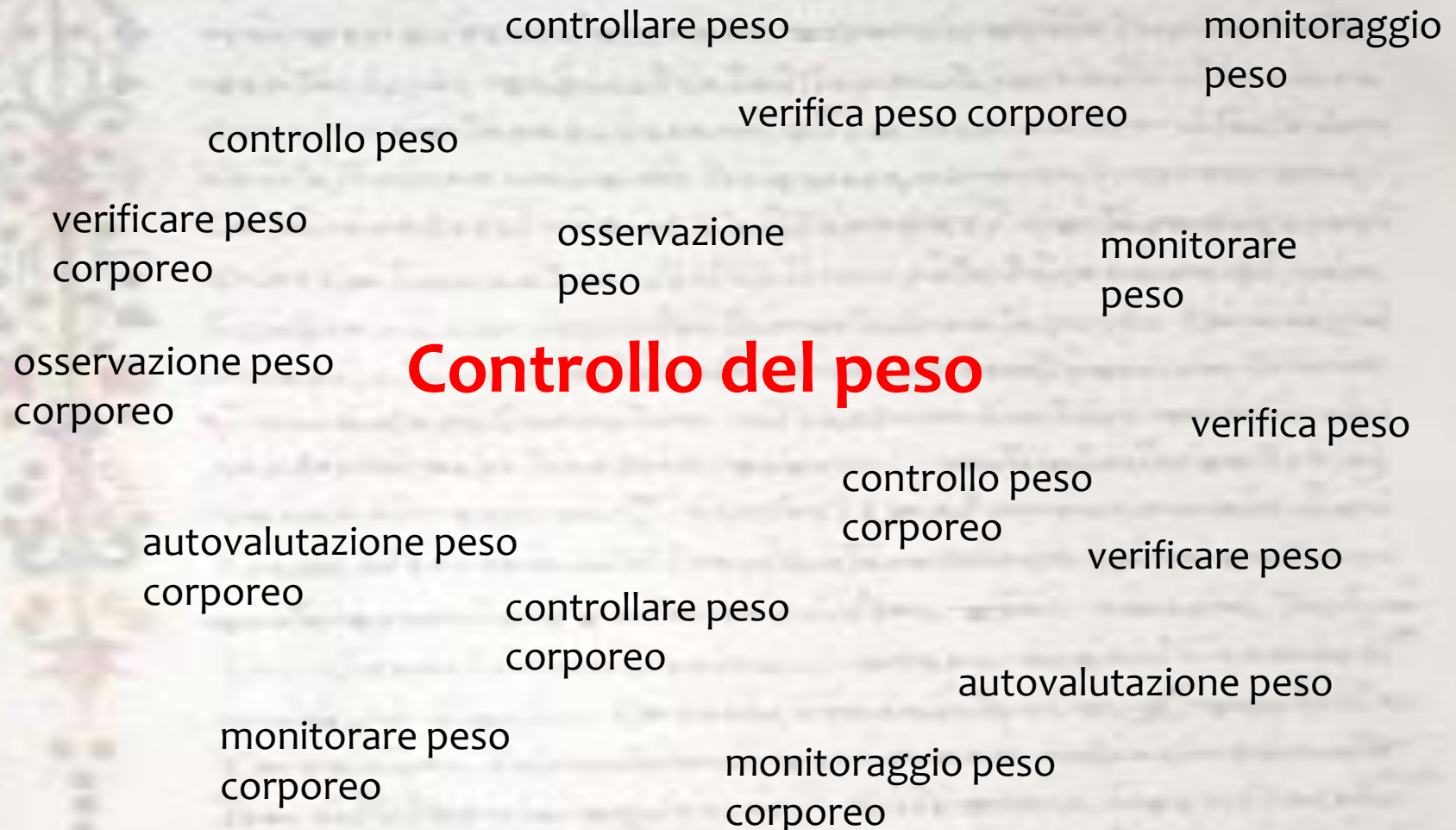
WORD	COUNT	PERCENT
FUROSEMIDE	409	1,492973
CONTROLLO	403	1,471071
COMPENSO	355	1,295857
INR	254	0,927176
TERAPIA	223	0,814017
NON	220	0,803066
QUADRO	218	0,795766
MIGLIORAMENTO	202	0,737361
RENALE	168	0,613251
FUNZIONE	162	0,591349
BISOPROLOLO	153	0,558496
SCOMPENSO	141	0,514692
LASIX	139	0,507392
CONTROLLI	138	0,503742



# Follow-up: un dizionario “empirico”

WORD	COUNT	PERCENT
CONTROLLO	403	1,471071
CONTROLLI	138	0,503742
RACCOMANDA	98	0,35773
REGIME	69	0,251871
PESO	67	0,24457
MONITORAGGIO	60	0,219018
CONSIGLIA	59	0,215368
PONDERALE	41	0,149662
RACCOMANDANO	28	0,102208
RISPARMIO	28	0,102208
ESEGUIRE	24	0,087607
SEGNALARE	24	0,087607
CORPOREO	22	0,080307
RIPOSO	22	0,080307

# Concetti e termini relativi



# L'esame del corpus: algoritmi per la ricerca

- **if** (controllo peso) or  
(controllare peso) or  
(monitoraggio peso) or  
(monitorare peso) or  
(autovalutazione peso) or  
(osservazione peso) or  
(verifica peso) or  
(verificare peso) or  
...  
(verificare peso corporeo)

**then** peso, controllo

- **else**  $\neg$  peso, controllo

Si ripete su tutti  
i testi del  
corpus, per tutti  
i concetti



# L'esame del corpus: alcuni risultati

Concetto	Presenza/assenza (campo "consigli")	Occorrenze	Percentuale
Riposo	0	402	97,34
Riposo	1	11	2,66
Dieta	0	406	98,31
Dieta	1	7	1,69
Alcool	0	411	99,52
Alcool	1	2	0,48
Fumo	0	408	98,79
Fumo	1	5	1,21
Controllo del peso	0	410	99,27
Controllo del peso	1	3	0,73



# Sviluppi...

- A breve termine
  - test set e **validazione**
    - valutare sensibilità e specificità del sistema
- A lungo termine
  - sviluppo delle **basi di conoscenze**
  - sviluppo delle **funzioni di nlp**

# Raccomandazioni conclusive

- Servono **competenze diverse**: mediche, informatiche, linguistiche, statistiche.
- Le **basi di conoscenze** sono decisive.
- Concetti e forme linguistiche: unire **teoria ed empiria**.

# Bibliografia (1/3)

- Feldman R. e Sanger J., *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, New York, NY, 2007
- Dulli S., Polpettini P. e Trotta M. (a cura di), *Text mining: teoria e applicazioni*, FrancoAngeli, Milano, 2004
- Cineca, *Text Mining: aspetti applicativi in campo biomedico*, Cineca, 2001
- Cohen A. M., Hersh W. R., *A survey of current work in biomedical text mining*, «Briefings in Bioinformatics», 2005, vol. 6, n. 1, pp. 57-71
- de Bruijn B., Martin J., *Getting to the core of knowledge: mining biomedical literature*, «International Journal of Medical Informatics», 2002, vol. 67, n. 1-3, pp. 7-18



# Bibliografia (2/3)

- Fiszman M., Chapman W., Aronsky D., Evans R. S., Haug P. J., *Automatic Detection of Acute Bacterial Pneumonia from Chest X-ray Reports*, «Journal of the American Medical Informatics Association», 2000, vol. 7, n. 6, pp. 593-604
- Forster A., Andrade J., van Walraven C., *Validation of a Discharge Summary Term Search Method to Detect Adverse Events*, «Journal of the American Medical Informatics Association», 2005, vol. 12, n. 2, pp. 200-206
- Hersh W., *Evaluation of biomedical text-mining systems: lessons learned from information retrieval*, «Briefings in bioinformatics», 2005, vol. 6, n. 4, pp. 344-356
- Hripcsak G., Austin J., Alderson P., Friedman C., *Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports*, «Radiology», 2002, vol. 224, n. 1, pp. 154-163



# Bibliografia (3/3)

- Melton G., Hripcsak G., *Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries*, «Journal of the American Medical Informatics Association», 2005, vol. 12, n. 4, pp. 448-457
- Murff H., Forster A., Peterson J., Fiskio J., Heiman H., Bates H., *Electronically Screening Discharge Summaries for Adverse Medical Events*, «Journal of the American Medical Informatics Association», 2003, vol. 10, n. 4, pp. 339-350
- Spasic I., Ananiadou S., McNaught J., Kumar A., *Text mining and ontologies in biomedicine: Making sense of raw text*, «Briefings in Bioinformatics», 2005, vol. 6, n. 3, pp. 239-251
- Swanson, D.R., *Complementary structures in disjoint science literatures*, in *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1991, ACM Press, Chicago, IL, pp. 280–289